Magyar főnévi WordNet-ontológia létrehozása automatikus módszerekkel

Miháltz Márton

MorphoLogic Kft.
1118 Budapest, Késmárki utca 8.
mihaltz@morphologic.hu

Kivonat. A cikk bemutatja a folyamatban lévő, magyar főnévi WordNet adatbázis létrehozását célul kitűző munkálatok módszereit és legfrissebb eredményeit. Bemutatjuk azt a 9 különböző számítógépes módszert, melyek célja magyar főnevek automatizált hozzárendelése az angol nyelvű, 1.6-os verziójú WordNet synsetjeihez. A felhasznált magyar főnevek egy elektronikus magyar-angol kétnyelvű szótár szóanyagából származnak. A heurisztikus hozzárendelések támogatásához a kétnyelvű mellett az egynyelvű magyar Értelmező Kéziszótár számítógéppel feldolgozható anyagából nyertünk ki strukturális és szemantikai információkat. A különböző folyamatok eredményeinek pontosságát egy kézzel egyértelműsített etalon halmaz segítségével becsültük meg, majd a főnévi adatbázist a validált eredményhalmazok különböző szintű pontosságot meghaladó kombinációival állítottuk elő.

Kulcsszavak: WordNet-építés, Magyar Főnévi WordNet, automatikus szemantikai információ-szerzés

1 Bevezetés

Napjainkban az intelligens számítógépes nyelvészeti alkalmazások—természetes nyelvi szövegek gépi feldolgozását segítő eszközök, keresőmotorok, fordítóprogramok—fejlesztésében egyre inkább szükség mutatkozik természetes nyelvi fogalomtárak, ontológiák, lexikális tudásbázisok alkalmazására. Az egyik legelterjedtebb nyelvi ontológiai formalizmus a WordNet (WN), mely eredetileg a mentális lexikon számítógépes modelljeként született ([5]). A WN a tartalmas szóosztályok (főnevek, igék, melléknevek és határozószók) lexikai elemeinek szemantikai hálózata, ahol a fogalmi csomópontokat *synsetek*, szinonima-halmazok alkotják, közöttük olyan szemantikai kapcsolatokkal, min például a hipernima („Az-egy") reláció.

A magyar WordNet létrehozását megcélzó projekt 2000-ben indult, első lépésként a legkiterjedtebb tartalmas szóosztály, a főnevek adatbázisának létrehozását megjelölve ([6]). Munkánk során automatikus eljárásokat alkalmaztunk az ún. kiterjesztési módszer („Extend method", [8]) megvalósítására, melynek lényege, hogy a szabadon hozzáférhető, angol nyelvű Princeton WordNet főnévi synsetjeit megfeleltetjük magyar főnevekkel. Mivel feltételezzük, hogy a főnévi fogalmak mind az angol, mind a magyar nyelvben hasonló szemantikai rendszerbe szerveződnek, hiszen ugyanazt a vi-

lágot írják le, ezzel a módszerrel gyorsan és hatékonyan előállítható egy magyar főnévi ontológia kiinduló változata, mely szemantikai kapcsolatait az angol WN-től örökli.

2 A felhasznált számítógépes erőforrások

Az alapfeladat (magyar főnevek angol synsetekhez kapcsolása) megvalósításának kiinduló anyaga a MorphoLogic Kft. angol-magyar szótári adatbázisa volt, mintegy 17 700 magyar főnévi címszóval, melyekhez 12 400, az angol WN által lefedett angol fordítás tartozik.

Az illesztési folyamat támogatására felhasznált másik erőforrás az XML formátumba konvertált *Magyar Értelmező Kéziszótár* (ÉKSz, [3]) anyaga volt. Az ÉKSz mintegy 42 000 főnévi címszót tartalmaz, melyekhez több mint 64 000 különböző szöveges definíció tartozik.

3 Az alkalmazott módszerek

Mivel a kétnyelvű szótár magyar címszavainak nagy része egynél több (az egész szótár anyagában átlagosan 1,7) angol fordítással rendelkezik, az angol megfelelők pedig az angol WN-ben gyakran többértelműek, azaz egynél több (átlag 2,16) synsethez tartoznak, a megfeleltetés során egyértelműsíteni kellett, vagyis a lehetséges angol synsetek közül kiválasztani azokat, amelyekhez a magyar szó tartozik (jelenti őket). Ezt a feladatot automatikus módon, 9 különböző—részben korábbi, hasonló projektek során ([1], [2]), részben általunk kifejlesztett—heurisztikus eljárás alkalmazásával oldottuk meg. Ezzel a módszerrel a költséges manuális munka csupán az eredmények ellenőrzésére redukálódik.

3.1 A kétnyelvű szótár anyagával támogatott módszerek

A heurisztikák első csoportja a kétnyelvű szótárból kinyerhető információkon alapszik. Ezek egy része, melyeket a eredetileg a spanyol WordNet létrehozásakor fejlesztettek ki Atserias és munkatársai ([1]), a kétnyelvű szótár magyar és angol szavai, illetve az angol címszavak és a WN megfelelő synsetjei közötti kapcsolatokból kinyerhető információkat hasznosítják. A következő heurisztikákat alkalmaztuk:

- EGYJELENTÉSŰ FORDÍTÁSOK: ha egy magyar szó valamelyik angol fordítása egyértelmű a WN-ben, vagyis csupán egyetlen synsetbe tartozik, akkor létrehozunk egy kapcsolatot a magyar szó és a synset között.
- VARIÁNSOK: ha egy WN synset kettő vagy több olyan angol szót tartalmaz, melyeknek csupán egyetlen magyar fordításuk van, és az ugyanaz a magyar szó, akkor a magyar szót hozzárendeljük a közös synsethez.
- METSZET MÓDSZER: a magyar szavakat hozzárendeli azokhoz a synsetekhez, amelyek legalább kettőt tartalmaznak a szó angol fordításai közül.

Egy negyedik, általunk kifejlesztett heurisztika a kétnyelvű szótár magyar oldalából kinyerhető morfo-szemantikai információkon alapul. A magyar címszavak egy része termékeny endocentrikus (főnév+főnév) szóösszetétel Az ilyen összetételű szavak egy részének jellemző tulajdonsága, hogy a szerkezet alaptagja—az összetétel utótagja—többnyire meghatározza azt a szemantikai mezőt, amelynek az összetétel által jelölt dolog eleme ([4]). Így például a *hangversenyzongora* összetett szót számítógépes morfológiai elemzéssel felbontva a *hangverseny+zongora* morfémákra, az utótag kiválasztásával megkapjuk az összetett szó DERIVÁCIÓS HIPERNÍMÁJÁT („a *hangversenyzongora* az egy (fajta) *zongora*"). Ez az információ a következő részben leírt, módosított fogalmi távolság formula segítségével felhasználható a lehetséges angol synsetek feletti egyértelműsítéshez.

3.2 Az Értelmező Kéziszótár anyagát hasznosító módszerek

A Magyar Értelmező Kéziszótár főnévi definícióit a Humor morfológiai elemzőprogrammal ([6]) dolgoztuk fel, majd ez elemzett szövegben morfo-szintaktikai mintázatok heurisztikus keresésével ismertettünk fel szemantikai relációkat. Ezáltal képesek voltunk a címszóhoz 53 300 főnévi definícióban hipernímákat, 10 500 definícióban szinonimákat, illetve további 826 definícióban holonímákat és 584 esetben meronímákat azonosítani. Ezeknek a szemantikai információknak egy részét az alábbi módszerekkel használtuk fel a magyar címszavak WN-hez képest történő egyértelműsítéséhez:

- SZINONIMÁK: a magyar címszó angol fordításaihoz tartozó synsetek közül azt választjuk ki, amely a legtöbbet tartalmazza a szinonima angol fordításai közül (de legalább kettőt).
- HIPERNÍMÁK: azokban az esetekben, ahol mind a magyar címszónak, mind a hozzá azonosított hiperníma szónak volt angol fordítása a kétnyelvű szótárban, az 1. Ábrán bemutatott, módosított fogalmi távolság formula alkalamazásával választottuk ki a megfelelő angol synsetet. Az eredeti formulát Atserias és munkatársai fejlesztették ki ([1]).

$$dist'(w_1, w_2) = \min_{\substack{c_{1i} \in w1 \\ c_{2j} \in w2 \\ depth(c_{1i}) < depth(c_{2j})}} |path(c_{1i}, c_{2j})|$$

Fig. 1. A módosított fogalmi távolság formulát magyar főnévek és hipernímáik angol fordításainak párjaira alkalmaztuk. A képlet azt a két WN synsetet adja vissza, amely a WN hipernímahálózatában a legközelebb helyezkedik el egymáshoz. A magyar címszót a mélyebben lévő (a címszóhoz tartozó) synsethez rendeljük

Egy harmadik heurisztika a mintegy 1 500 ÉKSz címszóhoz megtalálható LATIN megfelelőket használja fel. Ezek általában állat- és növényfajok, rendszertani kategóriák, betegségek stb. latin nevei, melyek az angol WordNetben is megtalálhatók, így a latint egyértelműsítő közvetítőnyelvként felhasználva vihetjük végbe a hozzárendeléseket.

A kétnyelvű és az értelmező szótáron alapuló módszerek eredményeit az 1. Táblázat ismerteti.

Table 1. A különböző illesztési módszerek eredményei: illesztett magyar főnevek és WN synsetek, valamint a közöttük létrejött kapcsolatok számai

Módszer	Magyar főnevek	WN 1.6 synsetek	Kapcsolatok
Egyértelmű fordítások	8 387	5 369	9 917
Metszet módszer	2 258	2 335	3 590
Variáns módszer	164	180	180
DerivHip + FT	1 869	1 857	2 119
ÉKSz szinonimák	927	707	995
ÉKSz hipernímák + FT	5 432	6 294	9 724
ÉKSz latin megfelelők	1 697	838	848

3.3 Módszerek a lefedettség további növelésére

Azoknál a magyar főneveknél, ahol a magyar-angol szótár nem tartalmazott angol fordítást az ÉKSz alapján hozzájuk azonosított hiperníma vagy szinonima szavakhoz, két további módszerrel jutottunk angol fordítással rendelkező (derivációs) hipernímákhoz.

Az első módszer a 3.1 részben ismertetett eljárással, illetve termékeny főnév-főnév képzések felismerésével (pl. *ruhadarab* $\Rightarrow$ *ruha*) keres derivációs hipernímákat a szinonimákhoz és hipernímákhoz. Mivel a hiperníma-reláció tranzitív, a címszó hipernímájának (vagy szinonimájának) hipernímája is hipernímája lesz a címszónak.

A második eljárás kikeresi az azonosított hiperníma (vagy szinonima) szót az ÉKSz címszavai között, és amennyiben az egyértelmű (egyetlen definíciója van csak, tehát nincs szükség a jelentések közötti egyértelműsítésre), az ahhoz azonosított hiperníma szót használja fel (ha az rendelkezik angol fordítással).

Ezzel a két módszerrel 9,2%-os emelkedést tudtunk elérni az automatikusan illesztett magyar főnevek lefedettségében. Az automatikus módszerek összesen 13 948 magyar főnevet rendeltek hozzá 12 085 angol WN synsethez, 22 169 kapcsolatot létrehozva.

4. Az eredmények validációja és egyesítése

A különböző módszerek eltérő megbízhatóságúak, különböző pontosságú eredményeket produkálnak. Ezek pontos ellenőrzéséhez a kétnyelvű szótár teljes magyar oldalának anyagából véletlenszerűen kiválasztottunk 400 főnevet, melyek az angol fordításaikon keresztül összesen 2 201 lehetséges WN synsethez tartoznak. A lehetséges kapcsolatokat kézzel egyértelműsítettük, kiválasztva azokat, amelyek fennállnak és kitörölve azokat, amelyek nem állnak fenn a magyar szavak és az angol synsetek között. Ezzel a módszerrel létrehoztunk egy etalon halmazt, melynek segítségével elvégezhető a részeredmények megbízhatóságának becslése.

Elsőként megvizsgáltuk a 9 automatikus módszer eredményeit. Minden heurisztika esetében megállapítottuk a heurisztika és az etalon halmaz által közösen lefedett magyar szavak halmazát, az ezekhez a heuriszitka, illetve az etalon által rendelt kapcsolatokat, valamint ezeknek a kapcsolathalmazoknak a metszetét. Két mérőszámmal jellemeztük egy adott heurisztika megbízhatóságát. A pontosság (precision) érték a metszet halmaz és a heurisztika által létrehozott kapcsolathalmaz, a fedés (recall) érték pedig a metszet és az etalonban található kapcsolatok arányát jelzi. Az eredmények a 2. Táblázatban láthatók. Ebben azt is feltüntettük, hogy az adott módszer a kétnyelvű szótár teljes magyar oldalának milyen arányához rendelt kapcsolatokat (lefedettség (coverage) érték).

Table 2. Az etalon halmaz alapján számított pontosság és fedés értékek, valamint a kétnyelvű szótár magyar oldalának lefedettsége a különböző automatikus módszerek esetében, pontosság szerint csökkenő sorrendben. A latin ekvivalenseket felhasználó módszert nem tudtuk ezzel a módszerrel értékelni, mivel az jórészt szaknyelvi, az etalon halmaz általános szókincsében nem szereplő szavakhoz rendelt synseteket. Kézi mintavételezéssel és ellenőrzéssel ennek a módszernek a pontosságát kb. 80%-osra becsültük

Módszer	Pontosság	Fedés	Lefedettség
Variánsok	92.01%	50.00%	0.50%
Szinonimák	80.00%	39.44%	8.00%
DerivHip	70.31%	69.09%	17.50%
Lef. növ. 1.	67.65%	46.94%	7.50%
Egyért. ford.	65.15%	55.49%	69.25%
Metszet	58.56%	35.33%	17.50%
Lef. növ. 2.	58.06%	28.57%	6.00%
Hipernímák	48.55%	41.71%	49.25%

A különböző forrásokból származó eredmények egyesítésében a spanyol WN készítői által alkalmazotthoz hasonló módszert követtük ([1], [2]). Elsőként meghatároztunk két megbízhatósági küszöbértéket (70%, illetve 65%), majd egyesítettük azokat az eredményhalmazokat, amelyek az etalon halmaz segítségével végzett pontosságbecslések alapján elérték, vagy meghaladták ezeket a küszöbértékeket. Így létrejött két eredményhalmaz, körülbelül 70, illetve 65 százalékos becsült pontossággal.

Ezután létrehoztuk azoknak az eredmény-halmazoknak a páronkénti metszeteit, amelyek nem szerepeltek a fenti két halmazban, majd ezekre is elvégeztük a pontosságbecslést az etalon halmaz segítségével. A lehetséges 13 metszethalmaz közül 9 becsült pontosság-értéke lett 65 százalék, vagy annál magasabb (ebből 8 metszethalmaz 70% vagy magasabb becsült pontosságú). Ezeket a kiválasztott metszethalmazokat hozzáadtuk a két alaphalmazhoz, így tovább tudtuk növelni a lefedettséget anélkül, hogy a pontosság jelentősen csökkent volna. A dolog mögött az az elgondolás húzódik, hogy az alacsonyabb pontosságú módszerek is adhatnak a küszöbértéket meghaladó pontosságú eredményeket, amennyiben több külön forrás is megerősíti őket.

A két kiinduló halmaz, a kombinált metszethalmazok, valamint a két végleges halmaz adatai a 3. Táblázatban láthatók.

Table 3. A különböző eredmények kombinációiból előálló halmazokban található magyar szavak, angol synsetek és kapcsolataik száma, a halmazok becsült pontosságával

Eredményhalmaz	Szavak	Synsetek	Kapcsolatok	Pontosság
1. alaphalmaz	2 445	2 170	2 722	76,14%
További metszethalmazok	7 183	6 142	8 579	76,70%
1. végleges halmaz	**7 927**	**6 551**	**9 635**	**75,38%**
2. alaphalmaz	12 275	11 597	20 439	65,11%
További metszethalmazok	3 110	2 698	3 431	66,91%
2. végleges halmaz	**12 839**	**12 004**	**22 169**	**63,35%**

5. Összegzés, további munka

A magyar főnévi WordNet adatbázis kiinduló változatait különböző automatikus módszerek eredményeinek kombinációival állítottuk össze. Egy manuálisan létrehozott etalon halmaz segítségével becsült pontosságértékek alapján két, eltérő méretű és pontosságú halmazt hoztunk létre a további munka számára. A továbbiakban szeretnénk főként kézi munka alkalmazásával (a helytelen kapcsolatok kiszűrésével) növelni az eredmények megbízhatóságát, illetve tovább növelni a lefedett magyar szavak számát (a legpontosabbnak bizonyult heurisztikák alkalmazásával további kétnyelvű szótármodulokra).

Hivatkozások

1. Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997)
2. Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998)
3. Juhász, J., I., Szőke, G. O. Nagy, M. Kovalovszky (szerk.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest (1972)
4. Kiefer, F.: Jelentéselmélet. Corvina, Budapest (2001)
5. Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.
6. Prószéky, Gábor: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany (1996) 149–158
7. Prószéky, G. M. Miháltz: Automatism and User Interaction: Building a Hungarian WordNet. Proc. of the Third International Conference on Language Resources and Evaluation, Las Pa mas de Gran Canaria, Spain (2002)
8. Vossen, P.: Right or Wrong. Combining lexical resources in the EuroWordNet project. Proceedings of Euralex-96, Goetheborg (1996)